Prosper Loan Data Analysis

Homagni Bhattacharjee

Introduction

This dataset is related to the loan, borrowers, lenders, interest rates and stuffs like that. Prosper or Prosper Marketplace Inc. is a San Francisco, California based company specializing in loans at low interest rates to the borrowers.In this dataset, we are using the data from Posper to analyse it and trying to find the pattern in the Prosper data.

For the purpose of this analysis, we are using R, a high level programming language of the analysis with some of its most popular graphic package ggplot.

This data set contains 113,937 loans with 81 variables on each loan, including loan amount, borrower rate (or interest rate), current loan status, borrower income, borrower employment status, borrower credit history, and the latest payment information.

The data is available in the CSV format and can be downloaded from here. A detailed information of the variables is mentioned here

Out of these variables, the following were chosen for further analysis, which are as follows:

Term : Amount of month customers opted for loan

LoanStatus : Current status of the loan like chargedoff, completed, defauted etc. . .

EstimatedEffectiveYield : Yield of lenders from borrowers minus the processing fee and late fines

ProsperScore : Risk Factor score from 1 to 10. 10 being least risky

BorrowerAPR : The Borrower's Annual Percentage Rate (APR) for the loan.

BorrowerRate : The Borrower's interest rate for this loan.

ListingCategory..numeric. : Prosper rating for borrowers in numbers

EmploymentStatus : Current type of employment

Occupation : Occupation of borrower at the time of listing

EmploymentStatusDuration : How long the employee has been employed

StatedMonthlyIncome : Monthly income of the borrower

MonthlyLoanPayment : Monthly loan payment amount

LoanOriginalAmount : Original amount of the loan

LoanOriginationQuarter : Quarter of the month when loan was originated

Basic exploration of the datset

Now we will subset the identified variables from the original dataset and replace the Term levels from months to years.

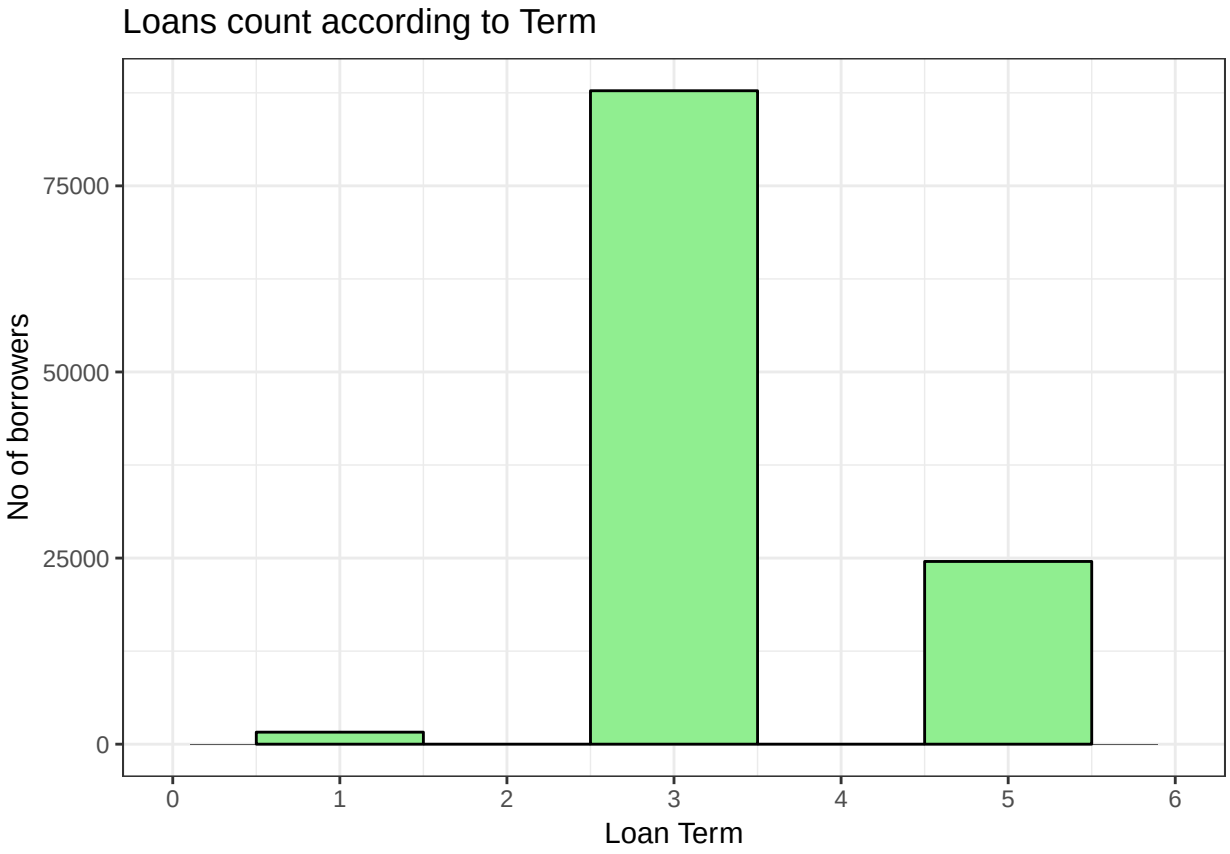
```
## 'data.frame':    113937 obs. of  81 variables:
##  $ ListingKey                       : Factor w/ 113066 levels "00003546482094282EF90E5",..: 7180 7
##  $ ListingNumber                    : int  193129 1209647 81716 658116 909464 1074836 750899 768193
##  $ ListingCreationDate              : Factor w/ 113064 levels "2005-11-09 20:44:28.847000000",..: 1
##  $ CreditGrade                      : Factor w/ 9 levels "","A","AA","B",..: 5 1 8 1 1 1 1 1 1 1 .
```

```
##  $ Term                                : int  36 36 36 36 36 60 36 36 36 36 ...
##  $ LoanStatus                          : Factor w/ 12 levels "Cancelled","Chargedoff",..: 3 4 3 4 4 4
##  $ ClosedDate                          : Factor w/ 2803 levels "","2005-11-25 00:00:00",..: 1138 1 12
##  $ BorrowerAPR                         : num  0.165 0.12 0.283 0.125 0.246 ...
##  $ BorrowerRate                        : num  0.158 0.092 0.275 0.0974 0.2085 ...
##  $ LenderYield                         : num  0.138 0.082 0.24 0.0874 0.1985 ...
##  $ EstimatedEffectiveYield             : num  NA 0.0796 NA 0.0849 0.1832 ...
##  $ EstimatedLoss                       : num  NA 0.0249 NA 0.0249 0.0925 ...
##  $ EstimatedReturn                     : num  NA 0.0547 NA 0.06 0.0907 ...
##  $ ProsperRating..numeric.             : int  NA 6 NA 6 3 5 2 4 7 7 ...
##  $ ProsperRating..Alpha.               : Factor w/ 8 levels "","A","AA","B",..: 1 2 1 2 6 4 7 5 3 3 .
##  $ ProsperScore                        : num  NA 7 NA 9 4 10 2 4 9 11 ...
##  $ ListingCategory..numeric.           : int  0 2 0 16 2 1 1 2 7 7 ...
##  $ BorrowerState                       : Factor w/ 52 levels "","AK","AL","AR",..: 7 7 12 12 25 34 18
##  $ Occupation                          : Factor w/ 68 levels "","Accountant/CPA",..: 37 43 37 52 21 43
##  $ EmploymentStatus                    : Factor w/ 9 levels "","Employed",..: 9 2 4 2 2 2 2 2 2 2 ...
##  $ EmploymentStatusDuration            : int  2 44 NA 113 44 82 172 103 269 269 ...
##  $ IsBorrowerHomeowner                 : Factor w/ 2 levels "False","True": 2 1 1 2 2 2 1 1 2 2 ...
##  $ CurrentlyInGroup                    : Factor w/ 2 levels "False","True": 2 1 2 1 1 1 1 1 1 1 ...
##  $ GroupKey                            : Factor w/ 707 levels "","00343376901312423168731",..: 1 1 335
##  $ DateCreditPulled                    : Factor w/ 112992 levels "2005-11-09 00:30:04.487000000",..: 1
##  $ CreditScoreRangeLower               : int  640 680 480 800 680 740 680 700 820 820 ...
##  $ CreditScoreRangeUpper               : int  659 699 499 819 699 759 699 719 839 839 ...
##  $ FirstRecordedCreditLine             : Factor w/ 11586 levels "","1947-08-24 00:00:00",..: 8639 661
##  $ CurrentCreditLines                  : int  5 14 NA 5 19 21 10 6 17 17 ...
##  $ OpenCreditLines                     : int  4 14 NA 5 19 17 7 6 16 16 ...
##  $ TotalCreditLinespast7years          : int  12 29 3 29 49 49 20 10 32 32 ...
##  $ OpenRevolvingAccounts               : int  1 13 0 7 6 13 6 5 12 12 ...
##  $ OpenRevolvingMonthlyPayment         : num  24 389 0 115 220 1410 214 101 219 219 ...
##  $ InquiriesLast6Months                : int  3 3 0 0 1 0 0 3 1 1 ...
##  $ TotalInquiries                      : num  3 5 1 1 9 2 0 16 6 6 ...
##  $ CurrentDelinquencies                : int  2 0 1 4 0 0 0 0 0 0 ...
##  $ AmountDelinquent                    : num  472 0 NA 10056 0 ...
##  $ DelinquenciesLast7Years             : int  4 0 0 14 0 0 0 0 0 0 ...
##  $ PublicRecordsLast10Years            : int  0 1 0 0 0 0 0 1 0 0 ...
##  $ PublicRecordsLast12Months           : int  0 0 NA 0 0 0 0 0 0 0 ...
##  $ RevolvingCreditBalance              : num  0 3989 NA 1444 6193 ...
##  $ BankcardUtilization                 : num  0 0.21 NA 0.04 0.81 0.39 0.72 0.13 0.11 0.11 ...
##  $ AvailableBankcardCredit             : num  1500 10266 NA 30754 695 ...
##  $ TotalTrades                         : num  11 29 NA 26 39 47 16 10 29 29 ...
##  $ TradesNeverDelinquent..percentage.  : num  0.81 1 NA 0.76 0.95 1 0.68 0.8 1 1 ...
##  $ TradesOpenedLast6Months             : num  0 2 NA 0 2 0 0 0 1 1 ...
##  $ DebtToIncomeRatio                   : num  0.17 0.18 0.06 0.15 0.26 0.36 0.27 0.24 0.25 0.25 ...
##  $ IncomeRange                         : Factor w/ 8 levels "$0","$1-24,999",..: 4 5 7 4 3 3 4 4 4 4
##  $ IncomeVerifiable                    : Factor w/ 2 levels "False","True": 2 2 2 2 2 2 2 2 2 2 ...
##  $ StatedMonthlyIncome                 : num  3083 6125 2083 2875 9583 ...
##  $ LoanKey                             : Factor w/ 113066 levels "00003683605746079487FF7",..: 100337
##  $ TotalProsperLoans                   : int  NA NA NA NA 1 NA NA NA NA NA ...
##  $ TotalProsperPaymentsBilled          : int  NA NA NA NA 11 NA NA NA NA NA ...
##  $ OnTimeProsperPayments               : int  NA NA NA NA 11 NA NA NA NA NA ...
##  $ ProsperPaymentsLessThanOneMonthLate: int  NA NA NA NA 0 NA NA NA NA NA ...
##  $ ProsperPaymentsOneMonthPlusLate     : int  NA NA NA NA 0 NA NA NA NA NA ...
##  $ ProsperPrincipalBorrowed            : num  NA NA NA NA 11000 NA NA NA NA NA ...
##  $ ProsperPrincipalOutstanding         : num  NA NA NA NA 9948 ...
```

```
##  $ ScorexChangeAtTimeOfListing        : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ LoanCurrentDaysDelinquent          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ LoanFirstDefaultedCycleNumber      : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ LoanMonthsSinceOrigination         : int  78 0 86 16 6 3 11 10 3 3 ...
##  $ LoanNumber                         : int  19141 134815 6466 77296 102670 123257 88353 90051 12126
##  $ LoanOriginalAmount                 : int  9425 10000 3001 10000 15000 15000 3000 10000 10000 1000
##  $ LoanOriginationDate                : Factor w/ 1873 levels "2005-11-15 00:00:00",..: 426 1866 260
##  $ LoanOriginationQuarter             : Factor w/ 33 levels "Q1 2006","Q1 2007",..: 18 8 2 32 24 33
##  $ MemberKey                          : Factor w/ 90831 levels "00003397697413387CAF966",..: 11071 1
##  $ MonthlyLoanPayment                 : num  330 319 123 321 564 ...
##  $ LP_CustomerPayments                : num  11396 0 4187 5143 2820 ...
##  $ LP_CustomerPrincipalPayments       : num  9425 0 3001 4091 1563 ...
##  $ LP_InterestandFees                 : num  1971 0 1186 1052 1257 ...
##  $ LP_ServiceFees                     : num  -133.2 0 -24.2 -108 -60.3 ...
##  $ LP_CollectionFees                  : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ LP_GrossPrincipalLoss              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ LP_NetPrincipalLoss                : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ LP_NonPrincipalRecoverypayments    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ PercentFunded                      : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ Recommendations                    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ InvestmentFromFriendsCount         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ InvestmentFromFriendsAmount        : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Investors                          : int  258 1 41 158 20 1 1 1 1 1 ...
```

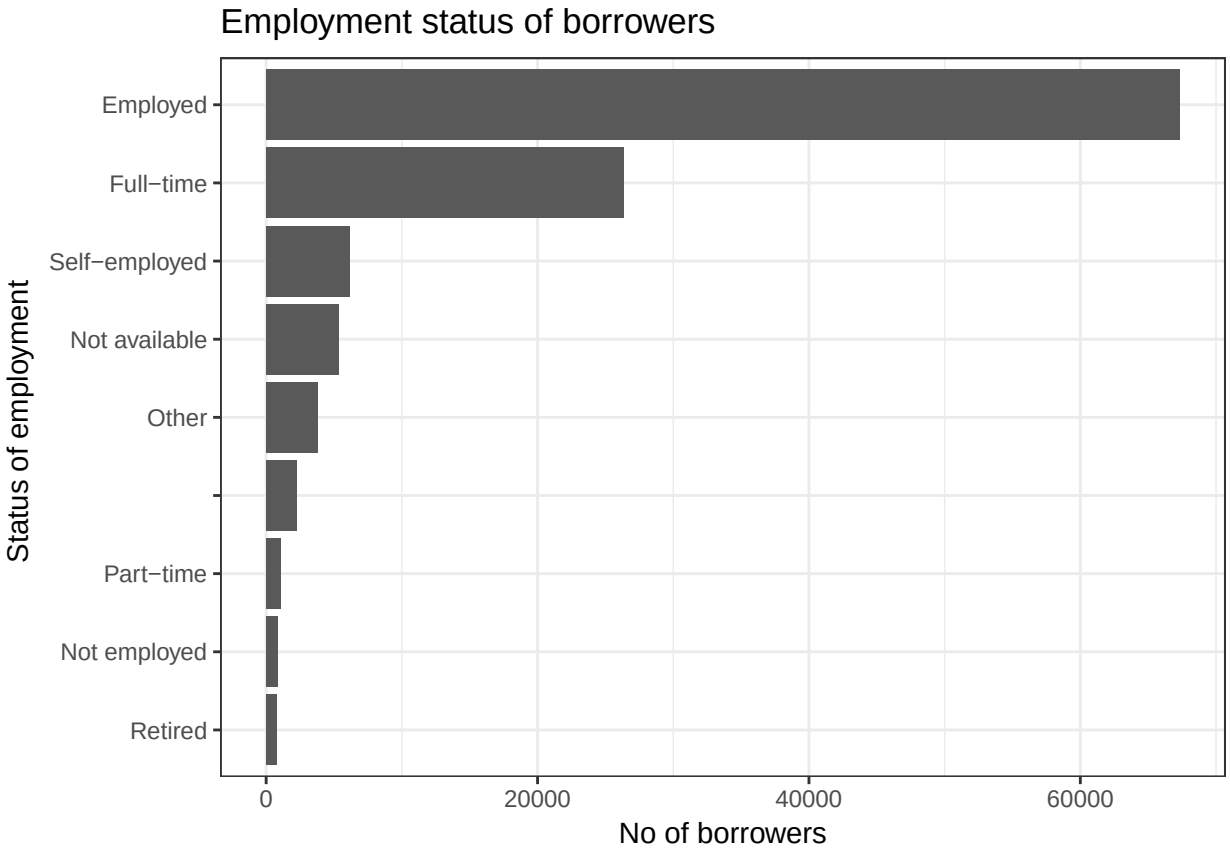
First we will look into the term of the loans.The first question that needs to be asked is **HOW LONG PEOPLE USUALLY OPT FOR LOAN?** Let's answer this question with a histogram

We can see that people don't really loan any amount for less than one year and the most popular loan amount is of 3 years although some people do choose for 5 years.



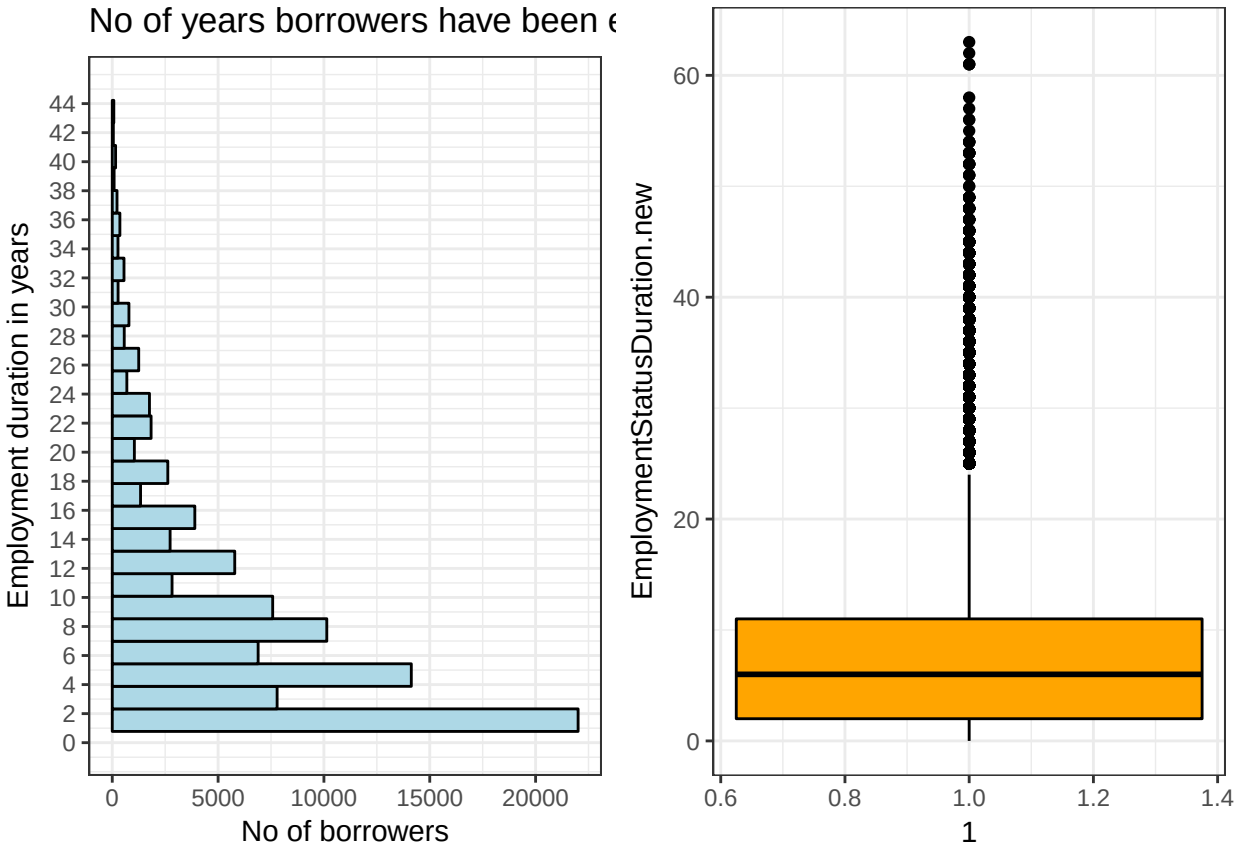
Employment Status

Now, let us take a look at the various employment status and occupation types that the borrowers have reported.



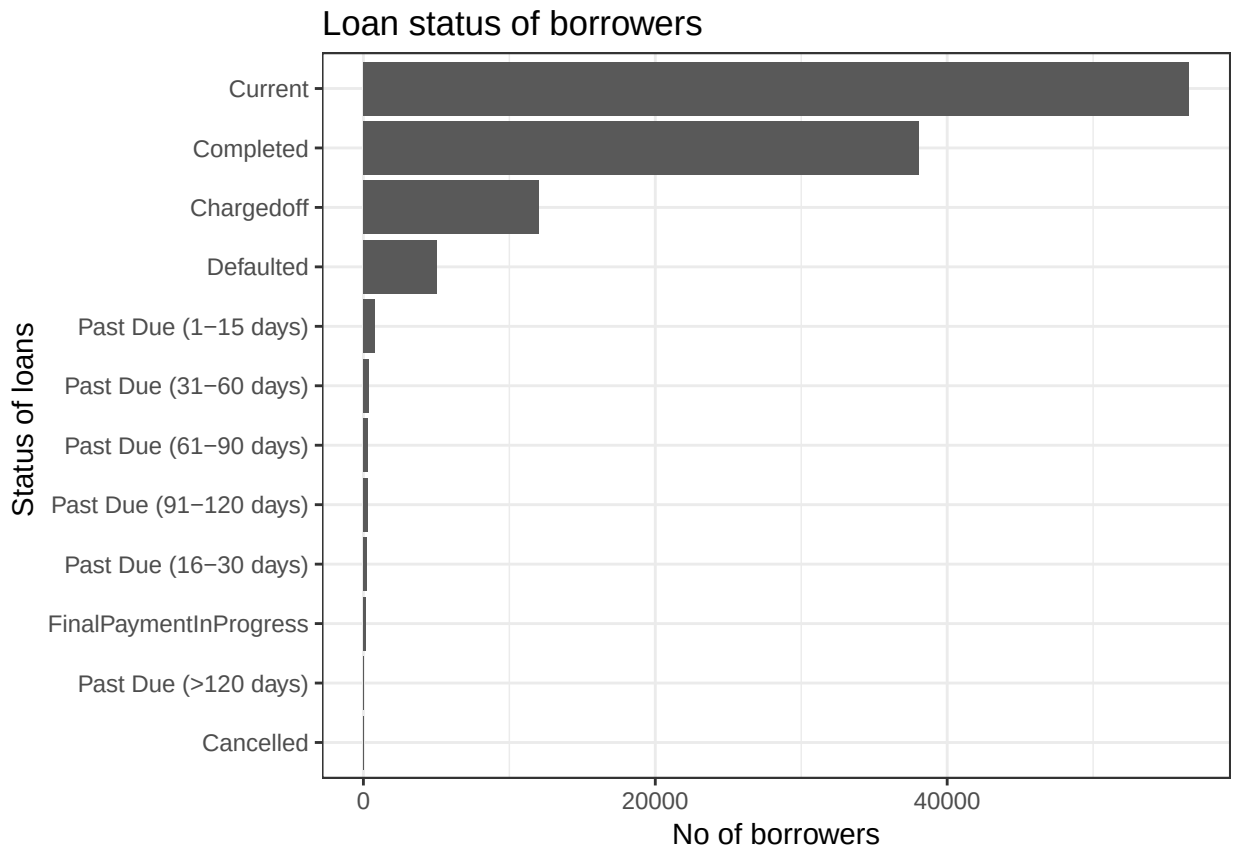
It is interesting to see the experience of the employed people and their relation with loans taken.

EmploymentStatusDuration

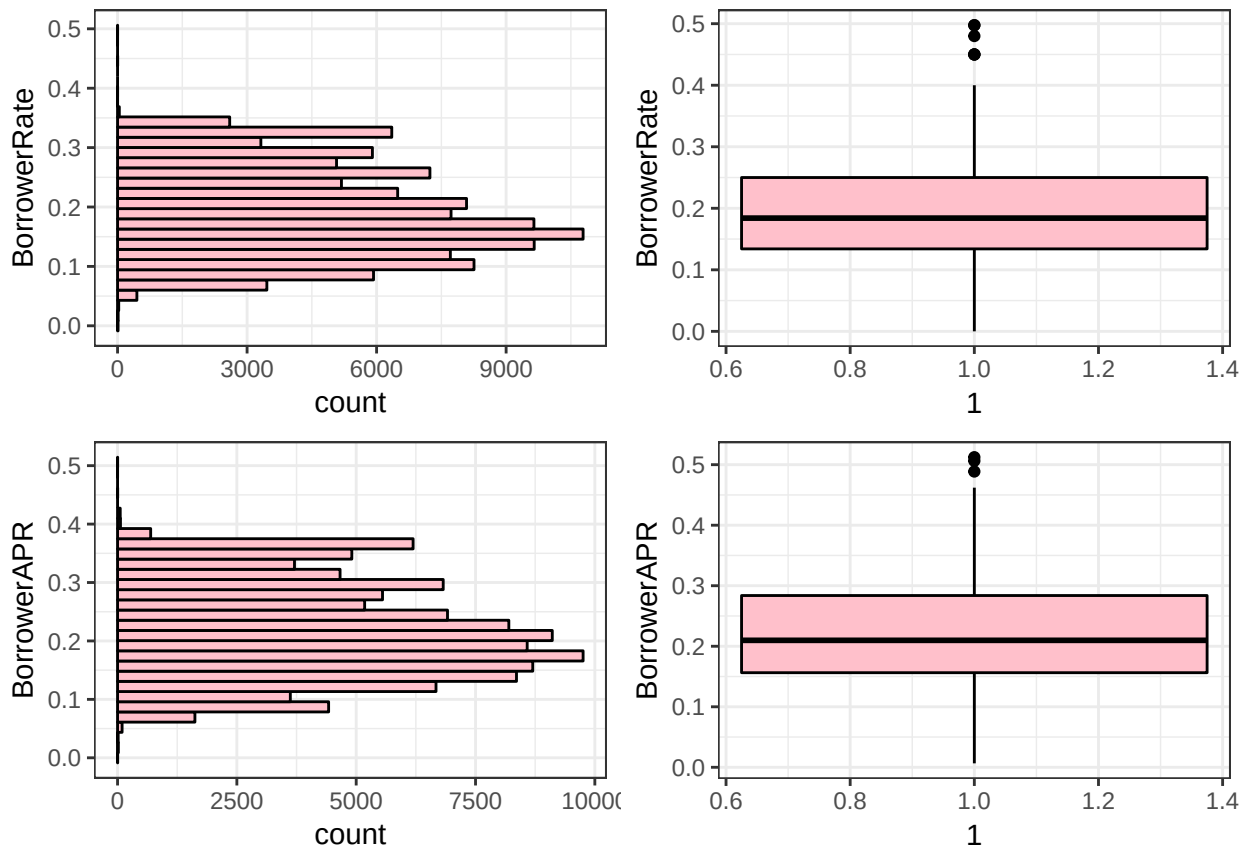


Clearly Employment status has a huge number of outliers. There is a high number of employees who are only recently starting out, hence the median value is low. Even though in the plot large number of outliers will remain, we will not remove it as it shows us that more experenced persons have low need for credit as they are more financially viable.

Loan Status


Loan status of borrowers
Current
Completed
Chargedoff
Defaulted
Past Due (1−15 days)
Past Due (31−60 days)
Past Due (61−90 days)
Past Due (91−120 days)
Past Due (16−30 days)
FinalPaymentInProgress
Past Due (>120 days)
Cancelled
Status of loans
0
20000
40000
No of borrowers

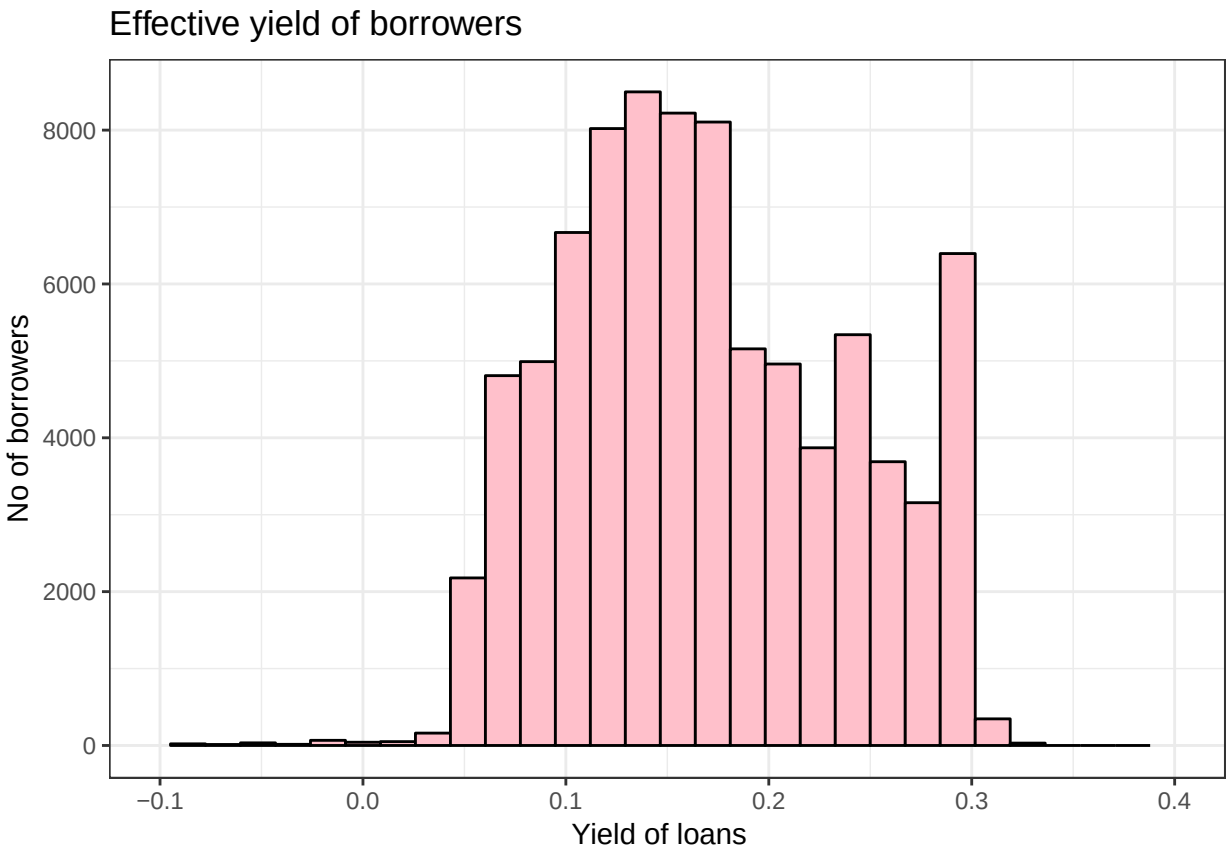
Borrower Rate & Borrower APR



Both these variables have very similar distribution. Very few outliers are present compared to the large number of observations and hence their effect may be ignored and dataset retained as it is.

Now from lenders perspective, they will be looking at **EstimatedEffectiveYield** as it is said to be better estimate for the lenders than the interest rate because the interest includes *processing fees*, *uncollected interest due to borrower being chargedoff*. Plus it also doesn't include *late fines*. Hence EstimatedEffectiveYield takes account for all these things and it is thus a better measure.

EstimatedEffectiveYield

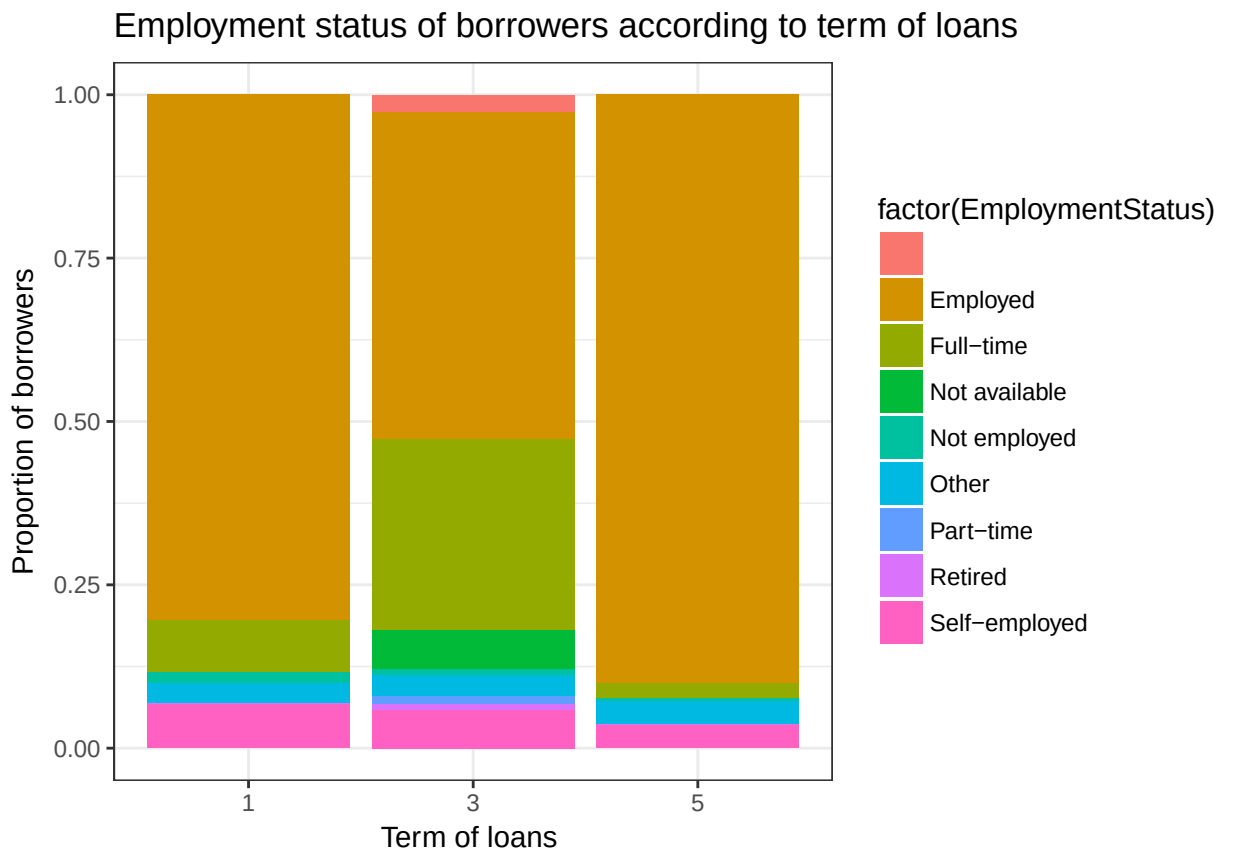


We can see that the EstimatedEffectiveYield is multimodal. We see the most popular EstimatedEffectiveYield is around 0.3 while the mean is around 0.17 represented by the blue dotted line. The multimodal pattern shows that there are multiple EstimatedEffectiveYield that is popular. Also, some customers have negative EstimatedEffectiveYield. This may mean a lot of things. This may mean that their BorrowerRate is a lot lower than their *service fee rate* or these customer's *uncollected interest on chargeoff* is lot more or they just never payed the late fee and payed back the loans along with the interest always on time.

Further analysis

We have seen that the borrowers took loans mostly of term 3 years. Let's assume something and check if it is correct or not. Here, I assume that people with better employment status take loans with longer terms and vice versa.

Now let us plot Term of loans with employment status of the borrowers



It is observed from the plot that the majority of the borrowers for all terms are a combination of (Employed + Full time) levels. This is could be because these borrowers are likely to have stable incomes which provide a level of financial stability. Also, lender may find them more loan worthy due to this factor.Hence it appears that the better employment status enables the borrowers to opt for longer term loans.

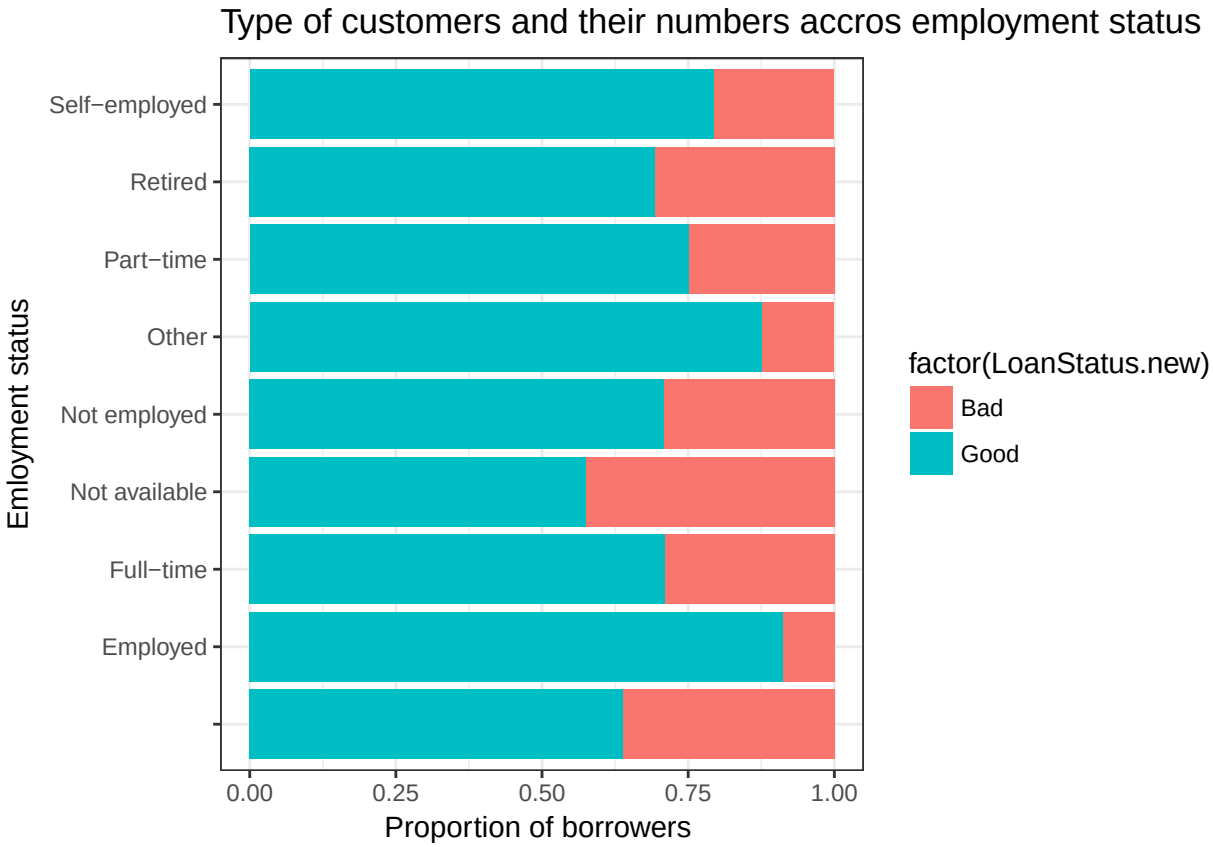
Now this leads to an interesting question, **WHETHER BETTER EMPLOYMENT STATUS MEANS BETTER BORROWERS** or not. For this, first let's see the distribution of LonaStatus variable.

For the purpose of this exploration, let us divide the borrowers into the following types

1. Good borrowers - Borrowers with loan status == "Current", "Completed", "FinalPaymentinProgress"

2. Bad borrowers- All other borrowers

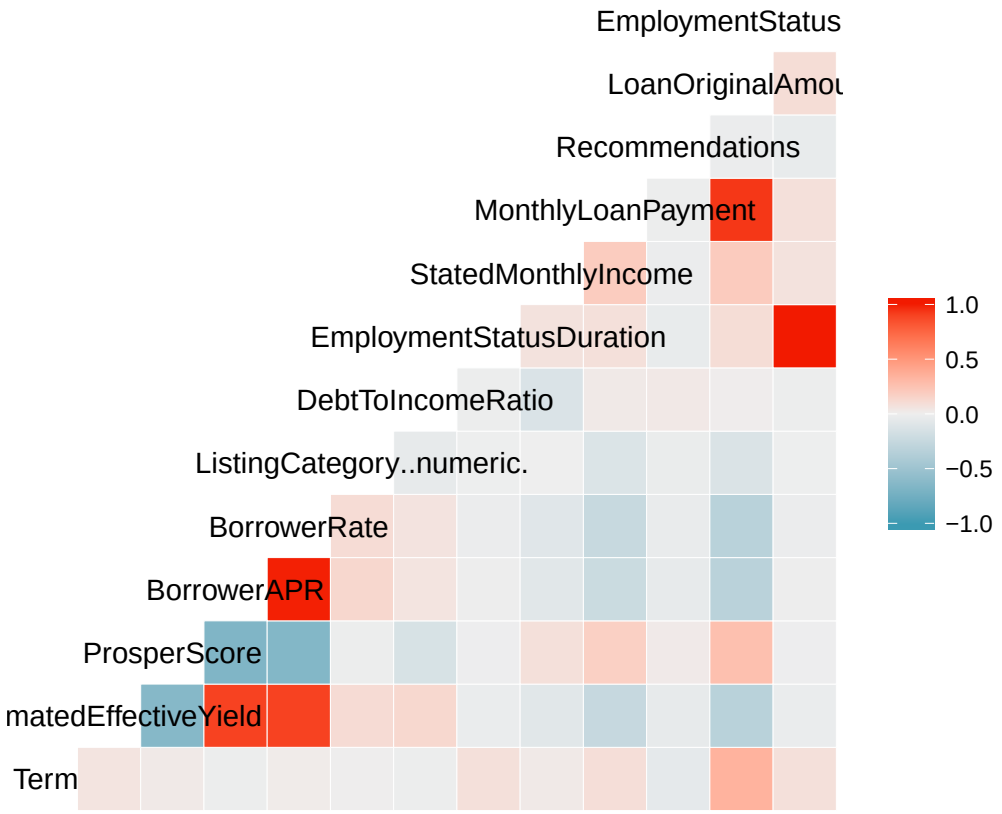
Now let us see the distribution of these two types of borrowers as a percentage among the various employment status categories.



It does appear from the plot that better employed people have better loan repayment status which may justify the lenders preferring such borrowers.

P.S. Some borrowers have undefined EmploymentStatus which might point out to fradulent loans being given out.

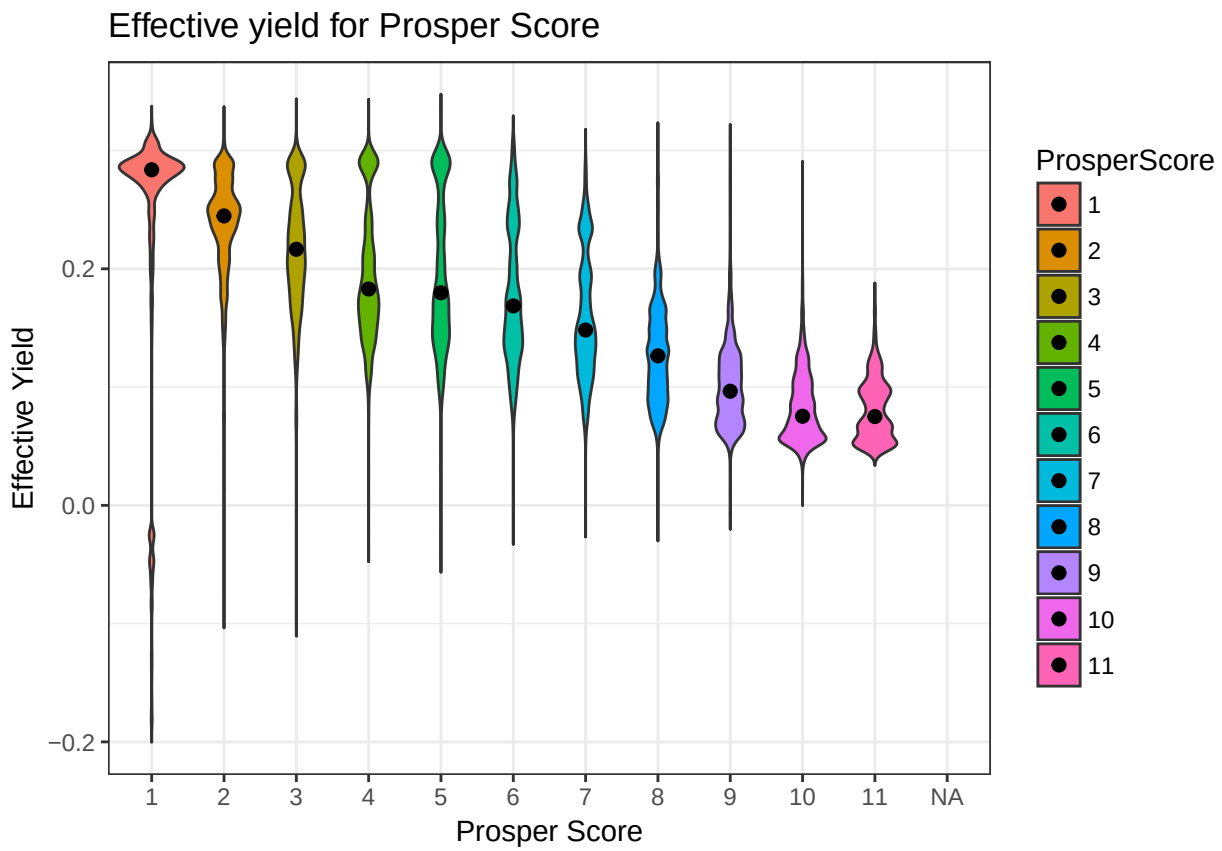
Looking into correlations



Do Lenders prefer borrowers with better Prosper Score ?

Now let's see what is the distribution of EstimatedEffectiveYield depending on the different **ProsperScore** which is a custom risk score built using historical Prosper data. The score ranges from 1-10, with 10 being the best, or lowest risk score.

This is important because we want to answer a question, i.e., **IF LENDERS GET MORE EstimatedEffectiveYield IF THEY HAVE BETTER ProsperScore?**



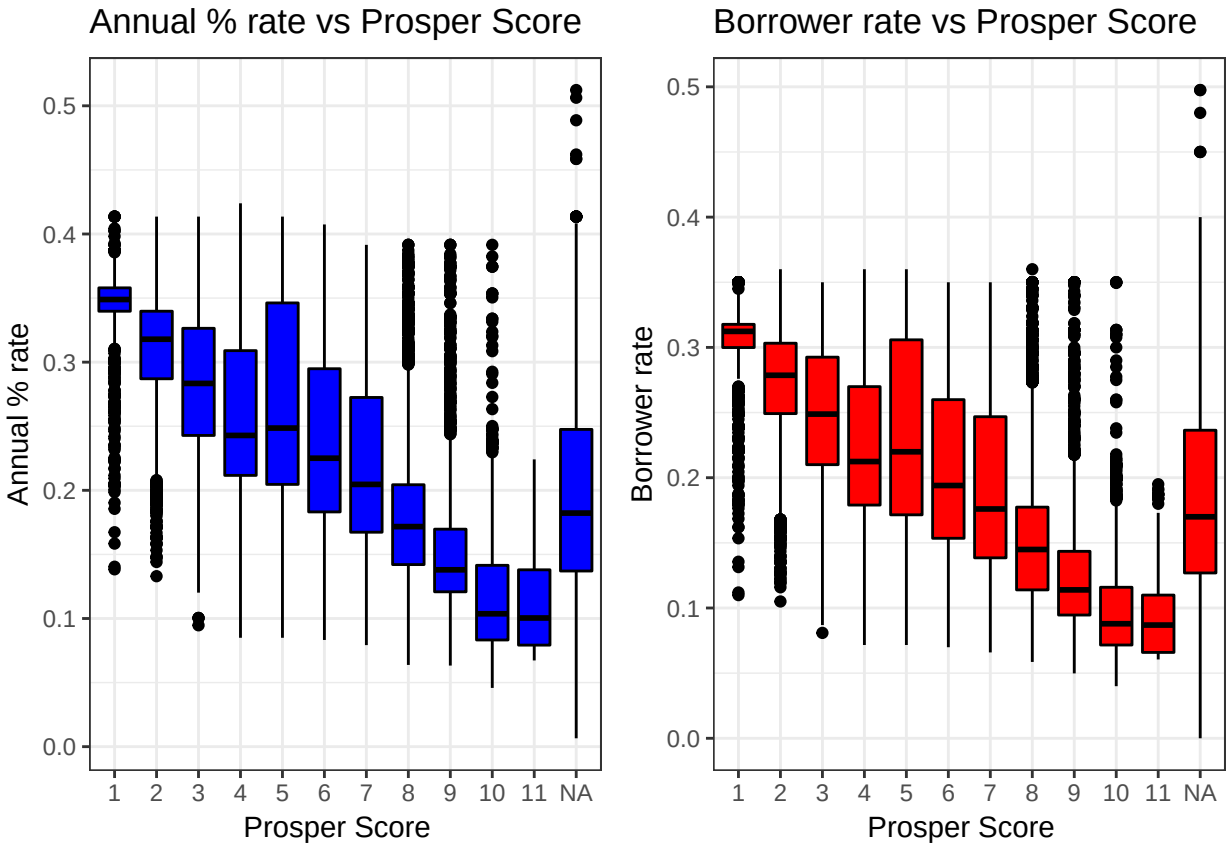
We can observe a trend here. Here more score for ProsperScore means better the borrower and lesser score means poor prospects from the borrowers.

We can see that for lower ProsperScore, distribution of effective yield in a lot more than the higher ProsperScore. This may mean that lenders charges a variety of interest rate from the borrower with poor prospects as compared to borrowers with better prospect. We can also notice how median (represented by the black dot) is decreasing as ProsperScore is increasing. This may mean that lenders give more relaxations to borrowers with better ratings as compared to borrowers with poor rating.

Does that mean lenders trust and like borrowers with better ProsperScore?

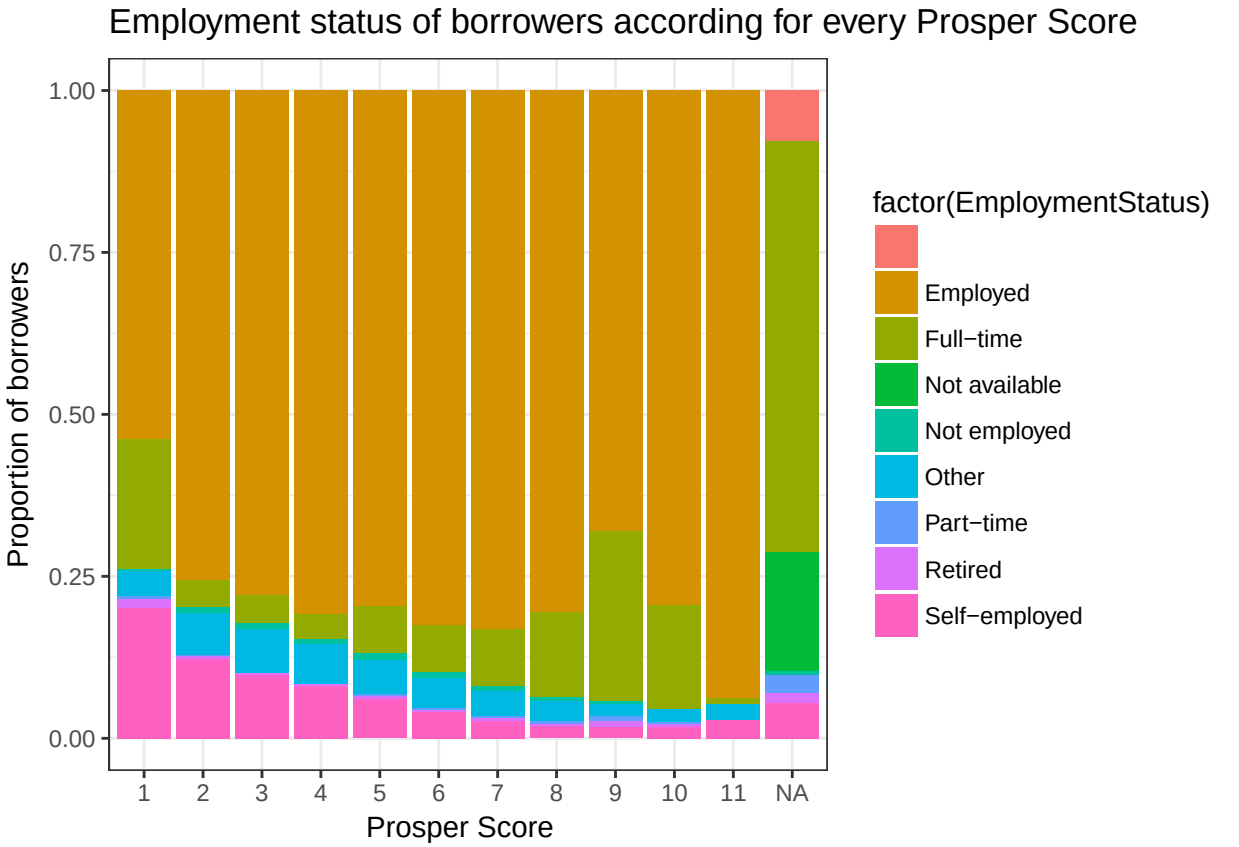
Let's do a little more analysis to reveal more. The reason we need more exploration on this is because EstimatedEffectiveYield includes more things such as late fine and doesn't include processing fee and others. So more EstimatedEffectiveYield for lesser ProsperScore borrowers may be due to high late fines because lesser ProsperScore borrowers are more prone to fail to repay their loan on time each month.

So, Let's see **if borrower's interest rate shows the same trend for each ProsperScore categories or not** because interest rates doesn't include late fines.



We can clearly observe that for both *BorrowerAPR* and *BorrowerRate* which are metric for interest rates, we see a declining trend as the *ProsperScore* is increasing. This shows that lenders charge less for all the borrowers with better ProsperScore as compared to borrowers with inferior ProsperScore.

Let us find out the prosper score for the different employment status.



Again, this is a confirmation that better employment status makes for better prosper score which indicates higher credit worthiness of these borrowers.

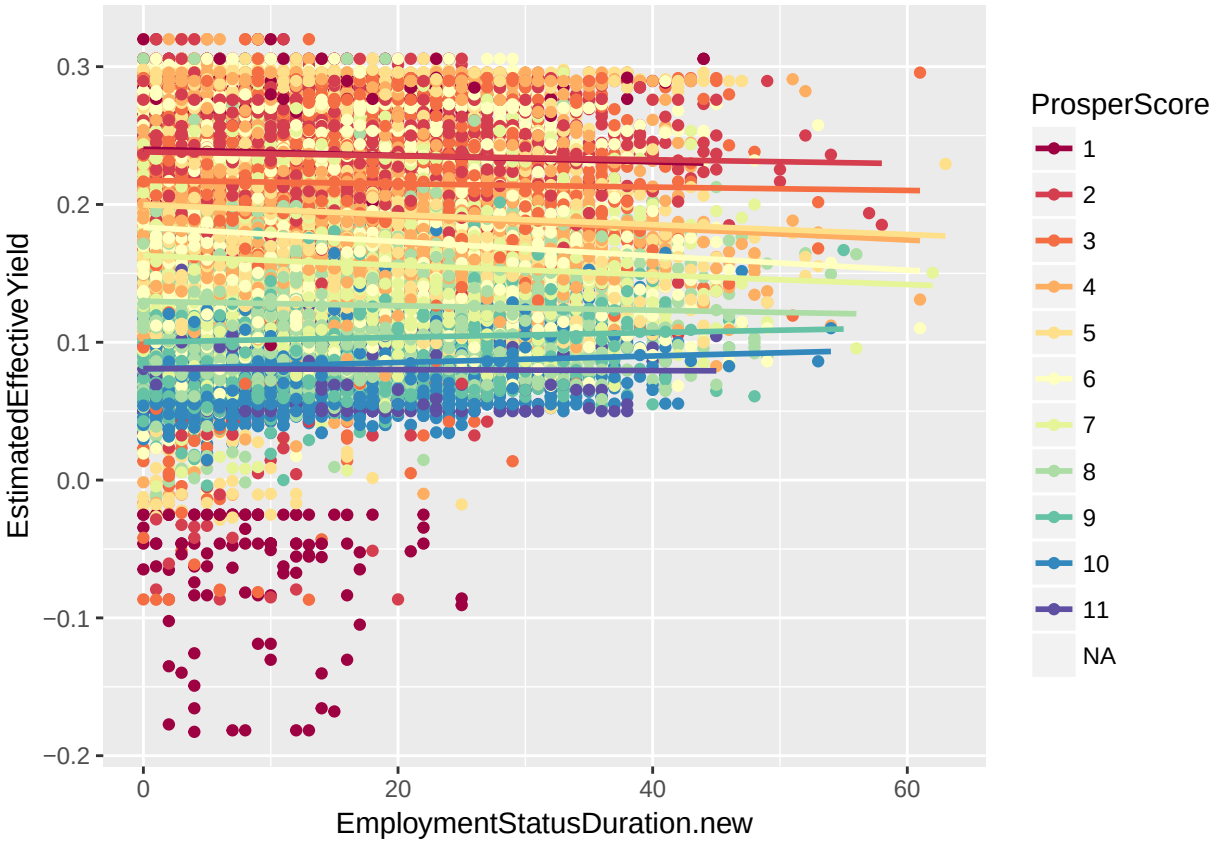
Looking at age of borrowers

We have seen before that young people tend to take more loans than their senior counterparts. This may be due the fact that as people gain experience their salary also increases and hence the lesser reason they find to opt for loans or the reason can be something different.

Lets explore it even more let's see the correlation between experience and the EstimatedEffectiveYield variable that we have explored earlier. The question that we want to explore -

DO LENDERS ASK FOR LESS INTEREST FORM THE BORROWERS WHO ARE MORE EXPERIENCED?

This can be true because people with more job experience should have more potential to replay their loan better because they have higher paying jobs and hence their ProsperScore would be higher. And as we have seen that borrowers with better prosper score pay lesser to the lenders and lenders somehow prefer them.



Correlation of Borrower's Experience As it seems from the scatter plot that the pattern seems to have no good correlation. It means our assumption was not correct. Borrowers with better EmploymentStatusDuration don't seem to get any special relaxation from lenders in terms of interest each month. This can be further confirmed by checking the Pearson's correlation Coefficient.

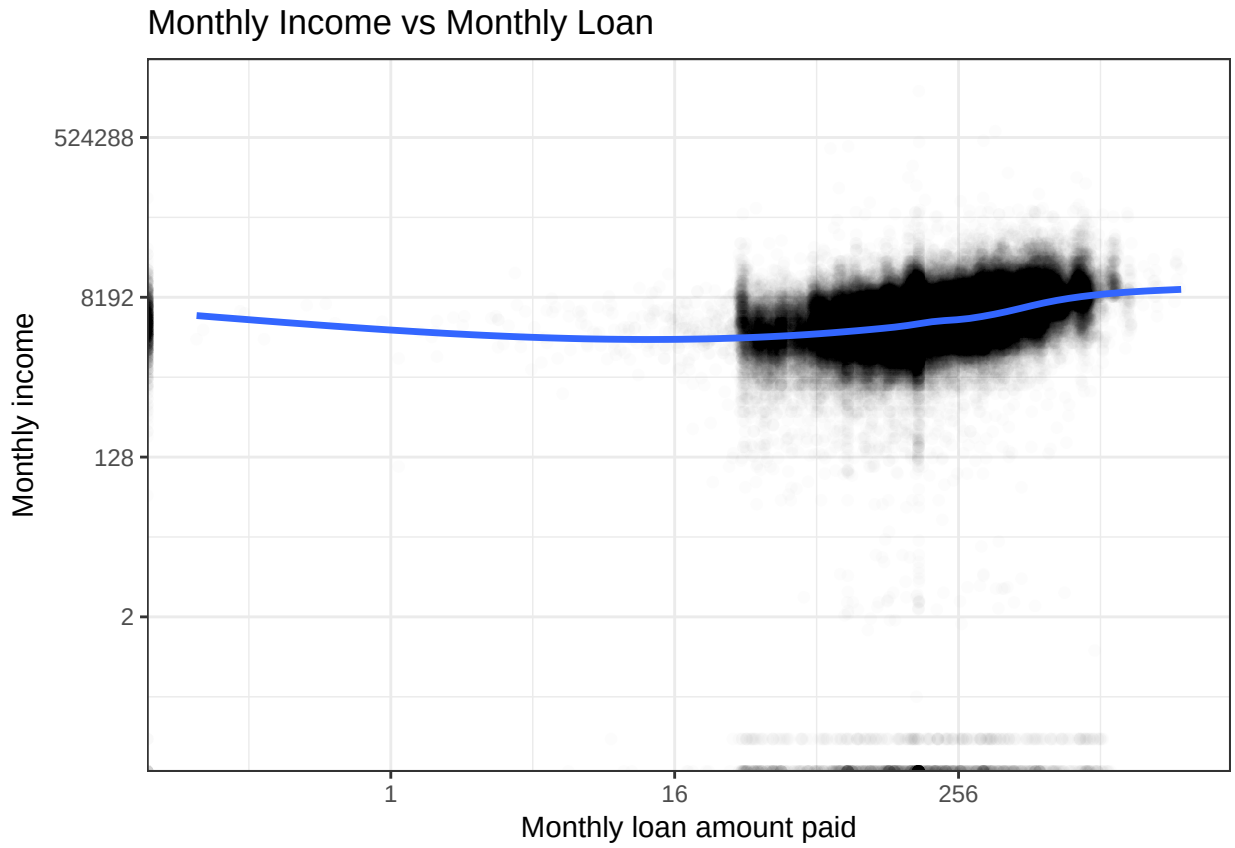
Calculating Pearson's product-moment correlation

```
##
##  Pearson's product-moment correlation
##
## data:  prosper_data_EDA$EmploymentStatusDuration.new and prosper_data_EDA$EstimatedEffectiveYield
## t = -6.7632, df = 84832, p-value = 1.359e-11
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.02993868 -0.01648751
## sample estimates:
##         cor
## -0.02321415
```

This also says that even though the true correlation is not true and alternative hypothesis is accepted, their is some serious statistical evidence of significance. But if we look into the CI, it is within the range of -0.03 to -0.016 which is very small. Good R value is said a value < -0.3 or value > 0.3. This value is definitely not that large. Judging from the context latest it is not. So we can say that there is no practical significance. Hence we can not tell with any confirmation that More Experienced Lenders end up paying Less/More interest to the Lenders.

Are Lenders greedy ?

I always wanted to do this. But from this dataset we can answer this question a lot of way. One of the way is to check if the lenders asked for money if the borrowers income was high. Let's see if the correlation is substantial.



MonthlyLoanPayment with *StatedMonthlyIncome* were plotted with both the scales are transformed in log scale. We can clearly that there is definite a strong positive correlation between monthly income and monthly loan amount.

Are we sure that they are greedy ?

Now we can see that there was definitely a strong correlation between the two variables but are we sure? Let's find the **Correlation Coefficient** to analyse it more.

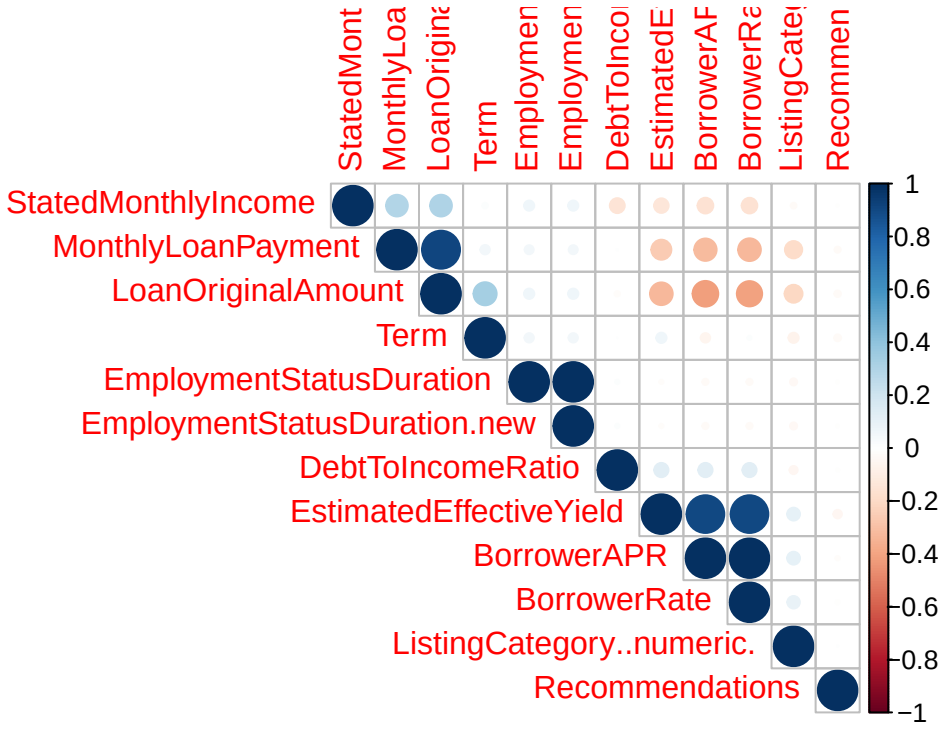
Calculating Pearson's product-moment correlation

```
##
##  Pearson's product-moment correlation
##
## data:  prosper_data_EDA$MonthlyLoanPayment and prosper_data_EDA$StatedMonthlyIncome
## t = 67.764, df = 113940, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1912423 0.2024055
```

```
## sample estimates:
##       cor
## 0.1968303
```

Well we can't really say that there is a strong correlation looking at the value of R which is almost 0.2. Usually it is said to be of high statistical importance if it is more than 0.3 or less than -0.3. But we can see that the value is still acceptable with somewhat positive correlation with the population Conficence Interval being more than 0. The strong t-statistics of 67.76 and small p-value shows that the statistical significance of alternative hypothesis is very strong.

More Correlations as a plot



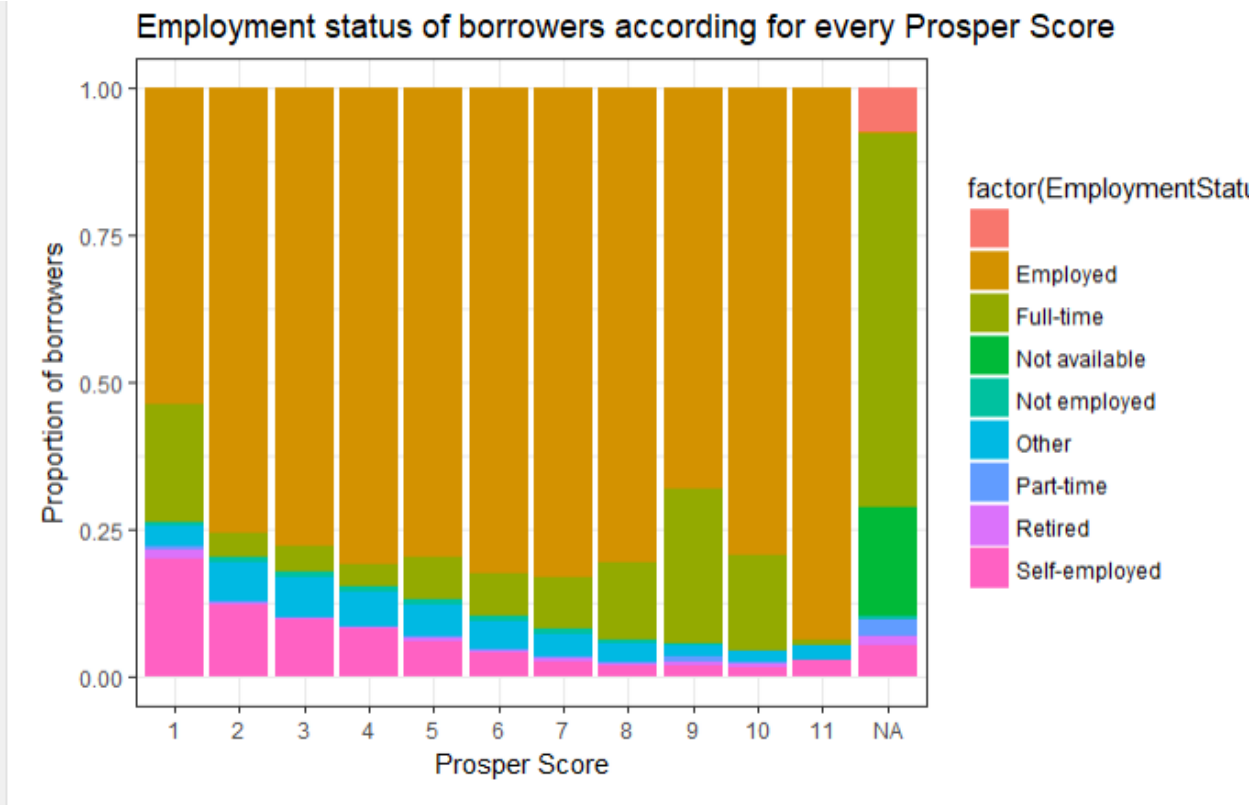
As we can see, *MonthlyLoanPayment* has strong correlation with *StatedMonthlyIncome* and *LoanOriginalAmount* and mild *EmploymentStatusDuration* and *Term*. This is not surprising as we have seen earlier that stable jobsallow for better financial condition to support loan payment.

Hence, for borrowers it would make more sense to track the Monthly income of the borrowers and the job types they hold to predict the monthly amount that the borrowers can repay.

Some Final Thoughts

Let's select 3 plots from what we have discussed and elaborate them bit further.

Good Job = Good borrower ?

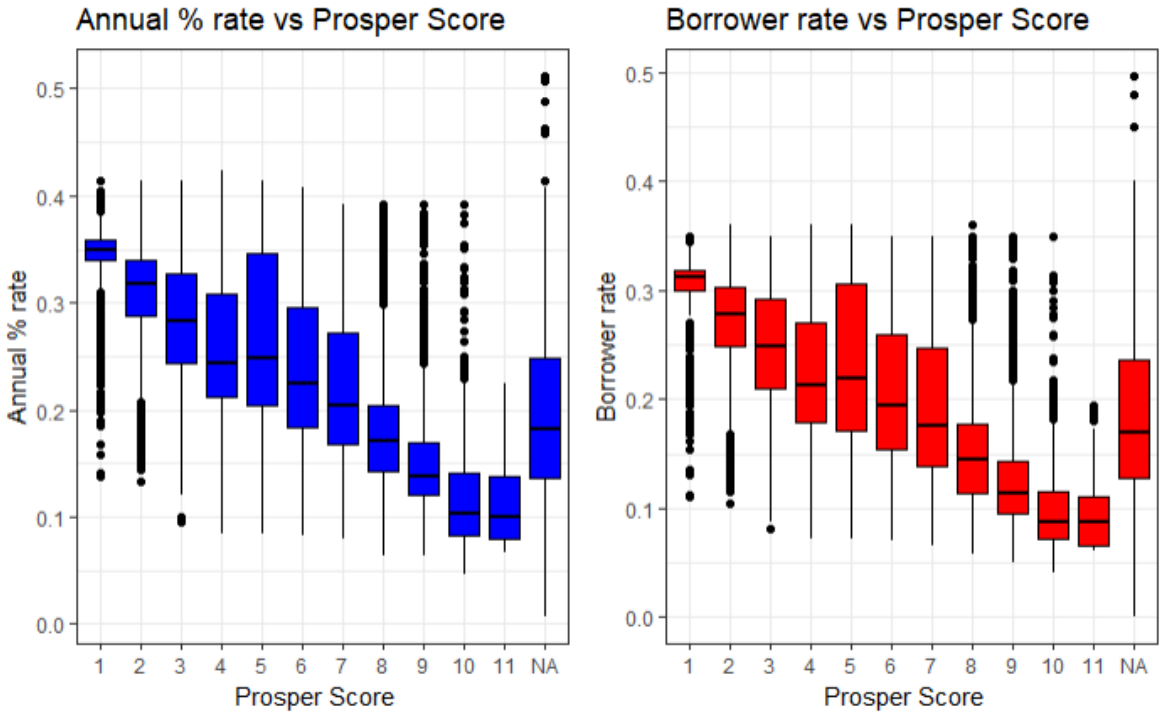


Now, we have seen that higher Prosper Score represents lower credit risk or better borrower. Now, we see that the higher scores have a trend of high % of (fulltime jobs + employed status). This indicates that the the lenders consider better job holders to have better repayment capacity and hence better borrowwers.

This assertion is borne out from the foolowing plot:

Here we can see that the better employed people have more % of borrowers with good loan status or good loan repayment status. (we may ignore others and NA category here as it is undefined). This gives credence to the aasertion.

Does lenders prefer borrowers with better ProsperScore ?



This depicts how we interest rates are affected by the Prosper Score for risk factor. As the score improves, the median interest rate shows a declining trend. This indeed proves that the lenders like to charge less from borrowers with better prosper score.

Reflection

1. Struggles:
   a. Large dataset which was difficult to process initially.
   b. Many variables like *BorrowerAPR*, *BorrowerRate*, *LenderYield*, *EstimatedEffectiveYield*. have similar meaning but not exactly the same. So it was difficult to choose right variables.
   c. FInding out the correct variable combination for the regression model.
2. Succeses:
   a. Managed to produce a good regression model which had good accuracy values.
3. Ideas for future:
   a. Regression model may be improve by using *Gradient Decend* to better approximate the slope and the intercept of the line.
   b. I also believe that is some information hidden inside the delinquency variables and late payment variables in *CurrentDelinquencies, AmountDelinquent, DelinquenciesLast7Years* etc with respect the Monthly loan amount which can be explored further.